


**METHOD AND DEVICE FOR SIMULTANEOUS VOICE RECOGNITION,
SPEAKER SEGMENTATION AND SPEAKER CLASSIFICATION**

Patent Number: JP2001060098

Publication date: 2001-03-06

Inventor(s): HAMEION SADARU MOHAMMAD BEIGI; TRITSCHLER ALAIN CHARLES LOUIS;
VISWANATHAN MAHESH

Applicant(s): INTERNATL BUSINESS MACH CORP

Requested Patent:  JP2001060098

Application
Number: JP20000188625 20000623

Priority Number
(s):

IPC Classification: G10L15/04; G06F3/16; G10L17/00; G10L15/00; G10L15/28

EC Classification:

Equivalents: CN1279462

Abstract

PROBLEM TO BE SOLVED: To obtain a method, in which audio information from an audio/video source is automatically transferred and a speaker is identified simultaneously, by transferring the audio source, simultaneously identifying latent segment boundaries and assigning a speaker label to each identified segment.

SOLUTION: The method includes a step, in which a transfer is made for an audio source to generate a text version of audio information, a step which simultaneously identifies latent segment boundaries, and a step in which a speaker label is assigned to each of identified segments. A simultaneous transfer, segmentation and speaker identification process 500 generates a transfer of audio information, which represents a speaker related to each segment, in real time. A segmentation process 600 identifies all frames in which segment boundaries may exist. A speaker identifying process 700 assigns a speaker label to each of the segments that use registered speaker databases.

Data supplied from the esp@cenet database - I2

(19) 日本国特許庁 (J P)

(12) 公開特許公報 (A)

(11) 特許出願公開番号

特開2001-60098

(P2001-60098A)

(43) 公開日 平成13年3月6日 (2001.3.6)

(51) Int.Cl. ⁷	識別記号	F I	テーマコード* (参考)
G 1 0 L 15/04		G 1 0 L 3/00	5 1 5 C
G 0 6 F 3/16	3 2 0	G 0 6 F 3/16	3 2 0 G
G 1 0 L 17/00		G 1 0 L 3/00	5 4 5 A
15/00			5 5 1 P
15/28			

審査請求 有 請求項の数23 O L (全 15 頁)

(21) 出願番号 特願2000-188625 (P2000-188625)

(22) 出願日 平成12年6月23日 (2000.6.23)

(31) 優先権主張番号 09/345237

(32) 優先日 平成11年6月30日 (1999.6.30)

(33) 優先権主張国 米国 (U S)

(71) 出願人 390009531

インターナショナル・ビジネス・マシーンズ・コーポレーション

INTERNATIONAL BUSINESS MACHINES CORPORATION

アメリカ合衆国10504、ニューヨーク州アーモンク (番地なし)

(72) 発明者 ハメイオン・サダル・モハマド・ベイギ
アメリカ合衆国ニューヨーク州、ヨークタウン・ハイツ、エッジヒル・ロード 3616

(74) 代理人 100086243

弁理士 坂口 博 (外2名)

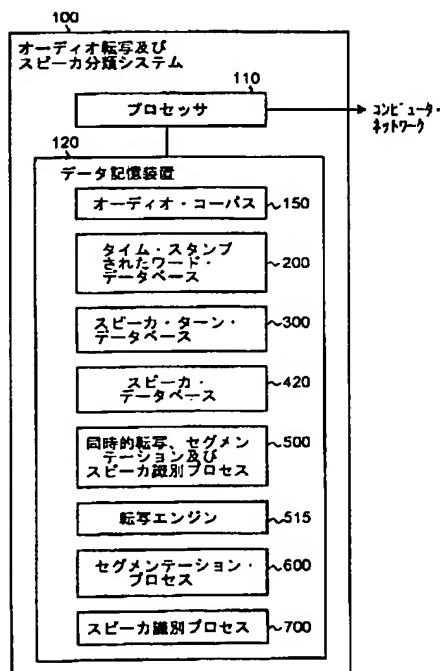
最終頁に続く

(54) 【発明の名称】 同時的な音声認識、スピーカ・セグメンテーション及びスピーカ分類のための方法及び装置

(57) 【要約】 (修正有)

【課題】 オーディオ／ビデオ・ソースからオーディオ情報を自動的に転写し、同時にスピーカを識別する方法及び装置。

【解決手段】 オーディオ転写及びスピーカ分類システムは音声認識システム、スピーカ・セグメンテーション・システム及びスピーカ識別システムを含む。音声認識システムは、各ワードに対してタイム・アライメントを伴う転写物を作成する。スピーカ・セグメンテーション・システムはスピーカを個別化し、非同種の音声部分相互間にセグメント境界が存在する可能性のあるすべてのフレームを識別する。スピーカ識別システムは、各識別されたセグメントにスピーカを割当てするため、登録済みのスピーカ・データベースを使用する。オーディオ／ビデオ・ソースからのオーディオ情報は、セグメント境界を識別するため同時に転写及びセグメント化された後、スピーカ識別システムは識別されたテキストの各部分にスピーカ・ラベルを割当てる。



【特許請求の範囲】

【請求項1】1つ又は複数のオーディオ・ソースからオーディオ情報を転写するための方法にして、前記オーディオ情報のテキスト・バージョンを作成するために前記オーディオ・ソースを転写するステップと、前記転写するステップと実質的に同時に前記オーディオ・ソースにおける潜在的なセグメント境界を識別するステップと、識別された各セグメントにスピーカ・ラベルを割り当てるステップと、を含む方法。

【請求項2】前記転写するステップは前記オーディオ・ソースにおける各ワードに対してタイム・アライメントを持った転写を作成することを特徴とする請求項1に記載の方法。

【請求項3】前記識別するステップは非同種の音声部分相互間にセグメント境界が存在するすべての可能なフレームを識別することを特徴とする請求項1に記載の方法。

【請求項4】前記割り当てるステップは登録されたスピーカ・データベースを利用してスピーカ・ラベルを各識別されたセグメントに割り当てることを特徴とする請求項1に記載の方法。

【請求項5】前記割り当てるステップは前記割り当てられたスピーカ・ラベルの信頼性を表すスコアを割り当てるステップを更に含むことを特徴とする請求項1に記載の方法。

【請求項6】前記割り当てるステップは前記割り当てられたスピーカ・ラベルに対して少なくとも1つの別の選択を割り当てるステップを更に含むことを特徴とする請求項1に記載の方法。

【請求項7】前記転写するステップ、識別するステップ、及び割り当てるステップはマルチ・スレッド環境では並列処理ブランチにおいて行われることを特徴とする請求項1に記載の方法。

【請求項8】前記識別するステップはBICモデル選択基準を使用してセグメント境界を識別することを特徴とする請求項1に記載の方法。

【請求項9】1つ又は複数のオーディオ・ソースからオーディオ情報を転写するための方法にして、前記オーディオ情報からフィーチャ・ベクトルを計算するステップと、

(a) 前記オーディオ・ソースを転写して前記オーディオ情報のテキスト・バージョンを作成するために、

(b) 前記オーディオ・ソースにおける潜在的なセグメント境界を識別するために、及び

(c) 各識別されたセグメントにスピーカ・ラベルを割り当てるために、前記フィーチャ・ベクトルを3つの並列処理ブランチに適用するステップと、を含む方法。

【請求項10】前記フィーチャ・ベクトルは共用メモリ・アーキテクチャを使用して前記並列処理ブランチに適

用されることを特徴とする請求項9に記載の方法。

【請求項11】前記共用メモリ・アーキテクチャは前記計算されたフィーチャ・ベクトルを前記並列処理ブランチの各々に対応するチャンネルに分配することを特徴とする請求項10に記載の方法。

【請求項12】前記転写するステップは前記オーディオ・ソースにおける各ワードに対してタイム・アライメントを持った転写物を作成することを特徴とする請求項9に記載の方法。

10 【請求項13】前記識別するステップは非同種の音声部分相互間にセグメント境界が存在するすべての可能なフレームを識別することを特徴とする請求項9に記載の方法。

【請求項14】前記割り当てるステップは登録されたスピーカ・データベースを利用してスピーカ・ラベルを各識別されたセグメントに割り当てることを特徴とする請求項9に記載の方法。

20 【請求項15】前記割り当てるステップは前記割り当てられたスピーカ・ラベルの信頼性を表すスコアを割り当てるステップを更に含むことを特徴とする請求項9に記載の方法。

【請求項16】前記割り当てるステップは前記割り当てられたスピーカ・ラベルに対して少なくとも1つの別の選択を割り当てるステップを更に含むことを特徴とする請求項9に記載の方法。

【請求項17】前記識別するステップはBICモデル選択基準を使用してセグメント境界を識別することを特徴とする請求項9に記載の方法。

30 【請求項18】1つ又は複数のオーディオ・ソースからオーディオ情報を転写するためのシステムにして、コンピュータ読み取り可能なコードを記憶するメモリと、

前記メモリに動作関係に結合され、前記コンピュータ読み取り可能なコードをインプリメントするように構成されたプロセッサと、

を含み、

前記コンピュータ読み取り可能なコードは、

前記オーディオ・ソースを転写して前記オーディオ情報のテキスト・バージョンを作成するように、

40 前記転写と実質的に同時に前記オーディオ・ソースにおける潜在的なセグメント境界を識別するように、及び各識別されたセグメントにスピーカ・ラベルを割り当てるように、

構成されることを特徴とするシステム。

【請求項19】コンピュータ読み取り可能なプログラム・コード手段を組み込まれたコンピュータ読み取り可能な媒体を含み、

前記コンピュータ読み取り可能なプログラム・コード手段は、

50 オーディオ情報のテキスト・バージョンを作成するため

にオーディオ・ソースを転写するステップと、
前記転写するステップと実質的に同時に前記オーディオ・ソースにおける潜在的なセグメント境界を識別するステップと、
識別された各セグメントにスピーカ・ラベルを割り当てるステップと、
を含むことを特徴とする製造物。

【請求項20】1つ又は複数のオーディオ・ソースからオーディオ情報を転写するためのシステムにして、コンピュータ読み取り可能なコードを記憶するメモリと、
前記メモリに動作関係に結合され、前記コンピュータ読み取り可能なコードをインプリメントするように構成されたプロセッサと、
を含み、
前記コンピュータ読み取り可能なコードは、
前記オーディオ情報からフィーチャ・ベクトルを計算し、

(i) 前記オーディオ・ソースを転写して前記オーディオ情報のテキスト・バージョンを作成するために、
(ii) 前記オーディオ・ソースにおける潜在的なセグメント境界を識別するために、及び
(iii) 各識別されたセグメントにスピーカ・ラベルを割り当てるために、前記フィーチャ・ベクトルを3つの並列処理ブランチに適用するように構成されることを特徴とするシステム。

【請求項21】コンピュータ読み取り可能なプログラム・コード手段を組み込まれたコンピュータ読み取り可能な媒体を含み、
前記コンピュータ読み取り可能なプログラム・コード手段は、
前記オーディオ情報からフィーチャ・ベクトルを計算するステップと、

(i) 前記オーディオ・ソースを転写して前記オーディオ情報のテキスト・バージョンを作成するために、
(ii) 前記オーディオ・ソースにおける潜在的なセグメント境界を識別するために、及び
(iii) 各識別されたセグメントにスピーカ・ラベルを割り当てるために、前記フィーチャ・ベクトルを3つの並列処理ブランチに適用するステップと、を含むことを特徴とする製造物。

【請求項22】1つ又は複数のオーディオ・ソースからオーディオ情報を転写するための方法にして、
前記オーディオ情報のテキスト・バージョンを作成するために前記オーディオ・ソースを転写するステップと、
前記オーディオ・ソースにおける潜在的なセグメント境界を識別するステップと、
識別された各セグメントにスピーカ・ラベルを割り当てるステップと、
前記転写するステップ、識別するステップ、及び割り当

てるステップと実質的に同時に前記テキスト・バージョンを前記割り当てられたスピーカ・ラベルと共に供給するステップと、
を含む方法。

【請求項23】1つ又は複数のオーディオ・ソースからオーディオ情報を転写するための方法にして、
前記オーディオ情報からフィーチャ・ベクトルを計算するステップと、

(i) 前記オーディオ・ソースを転写して前記オーディオ情報のテキスト・バージョンを作成するために、

(ii) 前記オーディオ・ソースにおける潜在的なセグメント境界を識別するために、及び

(iii) 各識別されたセグメントにスピーカ・ラベルを割り当てるために、前記フィーチャ・ベクトルを3つの並列処理ブランチに適用するステップと、
前記転写するステップ、識別するステップ、及び割り当てるステップと実質的に同時に前記テキスト・バージョンを前記割り当てられたスピーカ・ラベルと共に供給するステップと、

を含む方法。

【発明の詳細な説明】

【0001】

【発明の属する技術分野】本発明は、概して云えば、オーディオ情報分類システムに関し、詳しく云えば、オーディオ情報を転写(transcribe)し、オーディオ・ファイルにおけるスピーカ(発声者)を識別するための方法及び装置に関するものである。

【0002】

【従来の技術】放送ニュース機構及び情報検索サービスのような多くの機構は、記憶及び検索のために大量のオーディオ情報を処理しなければならない。オーディオ情報は、主題又はスピーカの名前、或いはそれらの両方によって分類されなければならないことが多い。主題によってオーディオ情報を分類するためには、先ず、音声認識システムが、自動分類又はインデキシングのために、オーディオ情報をテキストの形に転写する。しかる後、照会/ドキュメント・マッチングを行って関連ドキュメントをユーザに戻すためにインデックスが使用可能である。

【0003】従って、主題によってオーディオ情報を分類するというプロセスは本質的には完全に自動化されたものになっている。しかし、スピーカによってオーディオ情報を分類するというプロセスは、特に、放送ニュースのようなリアルタイムの応用に対しては、大きな労力を要する仕事を残すことが多い。スピーカ登録情報を使用してオーディオ・ソースからスピーカを自動的に識別するための数多くの計算主体のオフライン・テクニックが提案されているけれども、スピーカ分類プロセスはヒューマン・オペレータによって最も頻繁に行われ、ヒューマン・オペレータは各スピーカの変更を識別し、対応

するスピーカの識別を行う。

【0004】本発明の親出願（1999年4月9日出願の米国特許出願番号09/288,724号）は、オーディオ・コンテンツ（主題）及びスピーカのアイデンティティに基づいてオーディオ情報を検索するための方法及び装置を開示している。タイム・スタンプされたコンテンツ・インデックス・ファイル及びスピーカ・インデックス・ファイルを作成するために、インデキシング・システムがオーディオ情報を転写し、そしてインデックスする。しかる後、その生成されたコンテンツ及びスピーカ・インデックスは、オーディオ・コンテンツ及びスピーカ・アイデンティティに基づいて照会／ドキュメント・マッチングを行うために利用可能である。オーディオ・ソースからオーディオ情報を自動的に転写し、同時にスピーカをリアルタイムで識別する方法及び装置に対する要求が存在する。ベイズ情報基準（Bayesian Information Criterion-BIC）に基づいてスピーカ・セグメンテーション及びクラスタリングの改良を提供する方法及び装置に対する更なる要求も存在する。

【0005】

【発明が解決しようとする課題】従って、本発明の目的は、オーディオ／ビデオ・ソースからのオーディオ情報を自動的に転写し、同時にスピーカを識別するための方法及び装置を開示することにある。

【0006】

【課題を解決するための手段】開示されたオーディオ転写及びスピーカ分類システムは、音声認識システム、スピーカ・セグメンテーション・システム、及びスピーカ識別システムを含む。本発明の1つの局面によれば、オーディオ情報は、マルチスレッド環境における並列ブランチに沿って音声認識システム、スピーカ・セグメンテーション・システム、及びスピーカ識別システムによって処理される。

【0007】音声認識システムは、転写物を、その転写物内の各ワードに対するタイム・アライメントを伴って作成する。スピーカ・セグメンテーション・システムはスピーカを個別化し、非同種の音声部分相互間にセグメント境界が存在するすべての可能なフレームを識別する。しかる後、スピーカ識別システムは、登録されたスピーカ・データベースを使用して、各識別されたセグメントにスピーカを割り当てる。

【0008】本発明は、マルチスレッド環境における並列ブランチに沿って音声認識システム、スピーカ・セグメンテーション・システム、及びスピーカ識別システムによって処理されるフィーチャ・ベクトルを計算するために共通のフロント・エンド処理を利用する。一般に、フィーチャ・ベクトルは、例えば、計算されたフィーチャ・ベクトルを各チャネル（各処理スレッドに対応する）に分配するするためにサーバのような態様で作用する共用メモリ・アーキテクチャを使用して、3つの多重

処理スレッドに分配可能である。

【0009】本発明のもう1つの局面によれば、オーディオ／ビデオ・ソースからのオーディオ情報が同時に転写され及びセグメント境界を識別するためにセグメント化される。一旦音声セグメントがセグメンテーション・システムによって識別されると、スピーカ識別システムがその転写されたテキストの各部分にスピーカ・ラベルを割り当てる。

【0010】本願で開示されるセグメンテーション・プロセスは、オーディオ・データを通るパス上に、しかも、転写エンジンと同じパス上にあるセグメント境界であってスピーカ変更に対応するセグメント境界が存在するすべての可能なフレームを識別する。フレームは所定の期間にわたって音声特性を表す。セグメンテーション・プロセスは、2つのモデルを比較するモデル選択基準を使用して、所定のフレームにセグメント境界が存在するかどうかを決定する。第1モデルは、単一の全共分散ガウス分布（full-covariance Gaussian）を使用するサンプル（ x_1, \dots, x_n ）のウィンドウ内にセグメント境界が存在しないものと仮定する。第2モデルは、第1ガウス分布から得られた（ x_1, \dots, x_i ）及び第2ガウス分布から得られた（ x_{i+1}, \dots, x_n ）を持った2つの全共分散ガウス分布を使用するサンプル（ x_1, \dots, x_n ）のウィンドウ内にセグメント境界が存在するものと仮定する。

【0011】本願で開示されるスピーカ識別システムは、登録されたスピーカ・データベースを使用して各識別されたセグメントにスピーカ・ラベルを割り当てる。スピーカ識別プロセスはセグメンテーション・プロセスによって識別されたターンを、共用のフロント・エンドによって生成されたフィーチャ・ベクトルと共に受け取る。一般に、スピーカ識別システムは登録されたスピーカ・データベースにセグメント発声音（utterance）を比較し、「最も近似した」スピーカを見つける。そのスピーカ識別システムのためのモデル・ベース方式及びフレーム・ベース方式が開示される。

【0012】本発明の結果は、例えば、転写されたテキストを、割り当てられたスピーカ・ラベルと共に各セグメントに供給するユーザに直接に出力可能である。更に、本発明の結果は1つ又は複数のデータベースに記録可能であり、オーディオ・コンテンツ及びスピーカのアイデンティティに基づいてオーディオ情報に（及び間接的にはビデオに）参照を行うためにコンテンツ及びスピーカ・サーチ方法の結果を結合するという前記親出願において開示されたオーディオ検索システムのようなオーディオ検索システムによって利用可能である。

【0013】以下の詳細な説明及び図面を参照することによって、本発明の更に完全な理解及び本発明の更なる特徴及び利点の理解が得られるであろう。

【0014】

【発明の実施の形態】図1は、オーディオ/ビデオ・ソースからのオーディオ情報を自動的に転写し、同時にスピーカを識別するという本発明によるオーディオ転写及びスピーカ分類システム100を示す。オーディオ/ビデオ・ソース・ファイルは、例えば、オーディオ・レコーディングであってもよく、或いは、例えば、放送のニュース・プログラムからの生番組であってもよい。オーディオ/ビデオ・ソースは、先ず、転写され、同時に、スピーカの変更を表すセグメント境界が存在するすべての可能なフレームを識別するように処理される。

【0015】オーディオ転写及びスピーカ分類システム100は、音声認識システム、スピーカ・セグメンテーション・システム及びスピーカ識別システムを含む。音声認識システムは、転写物を、その転写物における各ワードに対するタイム・アライメントを伴って作成する。スピーカ・セグメンテーション・システムはスピーカを個別化し、セグメント境界が存在するすべての可能なフレームを識別する。セグメントは、所定のスピーカと関連したオーディオ・ソースの連続部分である。しかる後、スピーカ識別システムが各セグメントにスピーカ・ラベルを割り当てる。

【0016】図1は、本発明による例示的なオーディオ転写及びスピーカ分類システム100のアーキテクチャを示すブロック図である。オーディオ転写及びスピーカ分類システム100は、図1に示された汎用コンピュータ・システムのような汎用コンピュータ・システムとして具体化可能である。そのオーディオ転写及びスピーカ分類システム100はプロセッサ110及びデータ記憶装置120のような関連メモリを含む。なお、データ記憶装置120は分散型又はローカル型のものでよい。プロセッサ110は単一のプロセッサとして、又は並行して動作する複数のローカル・プロセッサ又は分散プロセッサとして実施可能である。データ記憶装置120及び/又は読取り専用メモリ（ROM）は1つ又は複数の命令を記憶するように動作可能であり、プロセッサ110はそれらの命令を検索、解釈、及び実行するように動作可能である。

【0017】望ましくは、データ記憶装置120は、本発明に従ってリアルタイムで処理可能な1つ又は複数の事前記録された又は生のオーディオ・ファイル又はビデオ・ファイル（或いは、それらの両方）を記憶するためのオーディオ・コーパス（corpus）データベース150を含む。又、データ記憶装置120は、図2に関連して後述するタイム・スタンプ・ワード・データベース200も含む。そのデータベース200は音声認識システムによって生成されたものであり、一組のタイム・スタンプされたワードを含む。図3に関連して後述するスピーカ・ターン・データベース300はスピーカ・セグメンテーション・システムと関連してスピーカ識別システムによって作成され、各セグメントの開始時間を、

1つ又は複数の対応する提案されたスピーカ・ラベルと共に表示する。図4と関連して後述するスピーカ・データベース420はスピーカ登録プロセス410によって作成され、各登録されたスピーカに対するエントリを含む。図1の例示的な実施例に示されたそれらの生成されたデータベース200及び300は、本発明の結果がリアルタイムでユーザに表示されるオンライン・インブリメンテーションに対しては必要とされず、その後のアクセスのためにも必要とされないことに注意してほしい。

【0018】更に、図5及び図6に関連して後述するように、データ記憶装置120は、同時転写、セグメンテーション及びスピーカ識別プロセス500、転写エンジン515、セグメンテーション・プロセス600、及びスピーカ識別プロセス700を含む。同時転写、セグメンテーション及びスピーカ識別プロセス500は転写エンジン515、セグメンテーション・プロセス600、及びスピーカ識別プロセス700の実行を調整する。同時転写、セグメンテーション及びスピーカ識別プロセス500はオーディオ・コーパス・データベース150における1つ又は複数のオーディオ・ファイルを分析し、各セグメントと関連するスピーカを表すオーディオ情報の転写をリアルタイムで作成する。セグメンテーション・プロセス600はスピーカを個別化し、セグメント境界が存在するすべての可能なフレームを識別する。スピーカ識別プロセス700は、登録されたスピーカ・データベースを使用する各セグメントにスピーカ・ラベルを割り当てる。

【0019】図2は、音声認識システムによって生成され、一組のタイム・スタンプされたワードを含む例示的なタイム・スタンプ・ワード・データベース200を示す。そのタイム・スタンプ・ワード・データベース200は、各々がその実施例における異なるワードと関連するレコード211乃至214のような複数のレコードを維持する。ワード・ストリング・フィールド220において識別された各ワードに対して、タイム・スタンプ・ワード・データベース200は開始時間フィールド230においてそのワードの開始時間を表示する。

【0020】図3は例示的なスピーカ・ターン・データベース300を示す。そのデータベース300は、スピーカ・セグメンテーション・システムと関連してスピーカ識別システムによって作成され、1つ又は複数の対応する提案されたスピーカ・ラベルと共に各セグメントの開始時間を表す。スピーカ・ターン・データベース300は、各々が実施例における種々のセグメントによって識別されるレコード305乃至308のような複数のレコードを維持する。フィールド320におけるセグメント番号によって識別された各セグメントに対して、スピーカ・ターン・データベース300は、オーディオ・ソース・ファイルの開始時間に関するそのセグメントの開始時間をフィールド330において表示する。更に、ス

ピーカ・ターン・データベース300は、フィールド340において各セグメントと関連するスピーカを、フィールド350における対応するスピーカ・スコアと共に識別する。1つのインプリメンテーションでは、スピーカ・ターン・データベース300はフィールド360において各セグメントと関連する1つ又は複数の代替スピーカ（次の最適な予測）を、フィールド370における対応する代替スピーカ・スコアと共に識別する。

【0021】A. スピーカ登録プロセス

図4はスピーカを登録又はエンロールするために使用される既知のプロセスを示す。図4に示されるように、各登録されたスピーカに対して、スピーカの名前が、パルス・コード変調（PCM）ファイルのようなスピーカ・トレーニング・ファイルと共にスピーカ登録プロセス410に供給される。スピーカ登録プロセス410はスピーカ・トレーニング・ファイルを分析し、スピーカ・データベース420において各スピーカに対するエントリを作成する。スピーカの音声サンプルをスピーカ・データベース420に加えるプロセスは登録と呼ばれる。その登録プロセスはオフラインであり、スピーカ識別システムは、関連するすべてのスピーカに対してそのようなデータベースが存在するものと仮定する。わずかな価値のオーディオに関して、一般には、各スピーカが複数のチャンネル及びマイクロフォンから複数の音響的条件を包含することを要求される。登録されたスピーカのトレーニング・データ又はデータベースは、それらのモデルへのアクセスが効率的な認識及び検索のために最適化されるように階層構造を使用して記憶される。

【0022】B. プロセス

前述のように、図5に示された同時転写、セグメンテーション及びスピーカ識別プロセス500は、転写エンジン515、セグメンテーション・プロセス600（図6）及びスピーカ識別プロセス700（図7）の実行を調整する。同時転写、セグメンテーション及びスピーカ識別プロセス500は、オーディオ・コーパス・データベース150における1つ又は複数のオーディオ・ファイルを分析し、各セグメントと関連するスピーカを表すオーディオ情報の転写をリアルタイムで作成する。図5に示されるように、同時転写、セグメンテーション及びスピーカ識別プロセス500は、まず、ステップ510においてオーディオ・ファイルからセブストラル（cepstral）フィーチャを既知の方法で抽出する。一般に、ステップ510はオーディオ信号のドメインを時間的ドメインから周波数ドメインに変更し、種々の周波数帯における信号エネルギーを分析し、その信号のドメインをセブストラル・ドメインに変更するためにもう1つの変換を使用する。

【0023】図5に示されるように、ステップ510は、転写エンジン515、セグメンテーション・プロセス600（図6）及びスピーカ識別プロセス700（図

7）に共通のフロント・エンド処理を提供する。一般に、ステップ510において計算されたフィーチャ・ベクトルは、転写エンジン515、セグメンテーション・プロセス（図6）及びスピーカ識別プロセス700（図7）に対応する3つの多重処理スレッドに分配可能である。それらのフィーチャ・ベクトルは、例えば、計算されたフィーチャ・ベクトルを各チャンネル（各処理スレッドに対応する）に分配するためにサーバのような態様で作用する共用メモリ・アーキテクチャを使用して3つの多重処理スレッドに分配可能である。

【0024】ステップ510において生成されたフィーチャ・ベクトルは、多重スレッド環境において並列ランチに沿って処理される。図5に示され且つ後述されるように、生成されたフィーチャ・ベクトルは多重スレッドを使用して

(i) ステップ515において転写エンジン、(ii) ステップ530において、図6に関連して後述されるスピーカ・セグメンテーション・プロセス600、及び(iii) ステップ560において、図7に関連して後述されるスピーカ識別プロセス700に適用される。

【0025】ステップ515において、それらの生成されたフィーチャ・ベクトルは、タイム・スタンプされたワードの転写ファイルを生成するために、IBM社から商業的に入手可能なViaVoice（商標）音声認識システムのような転写エンジンに供給される。しかる後、それらのタイム・スタンプされたワードは、ステップ520においてタイム・スタンプ・ワード・データベース200の中に任意選択的に収集可能である。更に、それらのタイム・スタンプされたワードは後述のステップ540においてインタリーバに供給される。

【0026】生成されたフィーチャ・ベクトルは、ステップ530において、図6に関連して後述されるセグメンテーション・プロセス600に適用される。一般に、セグメンテーション・プロセス600はスピーカを個別化し、非同種の音声部分相互間にセグメント境界が存在するすべての可能なフレームを識別する。セグメント境界が存在する各フレームはターンと呼ばれ、同種の各セグメントは単一のスピーカの音声に対応しなければならない。一旦セグメンテーション・プロセス600によって描出されると、各セグメントは（そのセグメントがスピーカ認識システムに対して要求される最小セグメント長の要件に合致すると仮定して）特定のスピーカによって発声されたものとして分類可能である。

【0027】セグメンテーション・プロセス600によって識別されたターンは、登録されたスピーカ・データベース420を使用して各セグメントにスピーカ・ラベルを割り当てるために、ステップ510において生成されたフィーチャ・ベクトルと共に、ステップ560において、図7に関連して後述されるスピーカ識別プロセス700に適用される。一般に、スピーカ識別システムは

セグメント発声音をスピーカ・データベース420に比較し(図4)、「最も近似した」スピーカを見つける。スピーカ識別プロセス700によって作成されたその割り当てられたスピーカ・ラベルは後述のステップ550に供給される。

【0028】ステップ515において転写エンジンによって作成されたタイム・スタンプ済みワードは、ステップ530においてセグメンテーション・プロセス600によって識別されたスピーカ・ターンと共に、ステップ540においてインタリーバに適用され、それらのターンをタイム・スタンプ済みワードとインタリーブさせ、切り離された音声セグメントを作成させる。しかる後、切り離された音声セグメント及びステップ560においてスピーカ識別システムにより生成されたスピーカ識別子がステップ550においてユーザに表示される。

【0029】1つのインプリメンテーションでは、切り離された音声セグメントは、それらがステップ540においてインタリーバによって作成された時にリアルタイムで表示される。更に、例示的な実施例では、そのスピーカ認識システムに対して要求される最小セグメント長は8秒である。従って、一般には、分離された音声セグメントの始まりが最初に与えられた後約8秒で、スピーカ識別ラベルがその転写されたテキストに付加される。切り離された音声セグメントがそのスピーカ認識システムに対して要求される最小セグメント長よりも短い場合、「未定(inconclusive)」のようなスピーカ・ラベルがそのセグメントに割り当て可能であることに注意すべきである。

【0030】C. ベイズ情報基準(BIC)の背景
前述のように、図6に示されたセグメンテーション・プロセス600はスピーカを個別化し、非同種の音声部分相互間にセグメント境界が存在するすべての可能なフレームを識別する。セグメント境界が存在する各フレームはターンと呼ばれ、同種の各セグメントは単一のスピーカの音声に対応しなければならない。一旦セグメンテーション・プロセス600によって描出されると、各セグメントは(そのセグメントがスピーカ認識システムに対して要求された最小セグメント長の要件に合致していると仮定して)特定のスピーカによって発声されたものとして分類可能である。セグメンテーション・プロセス600はベイズ情報基準(BIC)のモデル選択基準に基づくものである。BICは、 p 個のパラメータ・モデルのうちのどれが n 個のデータ・サンプル $x_1, \dots, x_n, x_i \in R^d$ を最もよく表すかを決定するために使用さ

れる漸近的に最適なベイズのモデル選択基準である。各モデル M_i は複数のパラメータ k_i を有する。サンプル x_i は独立したものであると仮定する。

【0031】BICの原理に関する詳細な検討のためには、例えば、The Annals of Statistics 誌の第6巻461-464ページ(1978)における G.Schwarz 氏による「モデルの寸法の見積り(Estimating the Dimension of a Model)」と題した論文を参照してほしい。そのBICの原理によれば、十分に大きい n に対して、データの最良のモデルは次式を最大化するものである。

$$BIC_i = \log L_i(x_1, \dots, x_n) - (\lambda k_i \log n) / 2$$

但し、 $\lambda = 1$ であり、 L_i はモデル M_i におけるデータの最大見込み値(換言すれば、 M_i の k_i パラメータに対する最大の見込み値をもったデータの見込み値)である。2つのモデルしか存在しない時、モデル選択のために簡単なテストが使用される。特に、 $\Delta BIC = BIC_1 - BIC_2$ が正である場合、モデル M_1 がモデル M_2 に優先して選択される。同様に、 $\Delta BIC = BIC_1 - BIC_2$ が負である場合、モデル M_2 がモデル M_1 に優先して選択される。

【0032】D. スピーカ・セグメンテーション
図6に示されたセグメンテーション・プロセス600は、セグメント境界が存在するすべての可能なフレームを識別する。汎用性を損なうことなく、精々1つのセグメント境界しか存在しない連続したデータ・サンプル (x_1, \dots, x_n) のウインドウを考察する。

【0033】フレーム i においてセグメント境界が存在するかどうかに関する基本的な疑問が次のような2つのモデル、即ち、モデル M_1 及びモデル M_2 の間のモデル選択問題として生じ得る。なお、モデル M_1 は (x_1, \dots, x_n) が単一の全共分散ガウス分布から得られる場合であり、モデル M_2 は (x_1, \dots, x_i) が第1ガウス分布から得られ、 (x_{i+1}, \dots, x_n) が第2ガウス分布から得られることによって (x_1, \dots, x_n) が2つの全共分散ガウス分布から得られる。

【0034】 $x_i \in R^d$ であるので、モデル M_1 は $k_1 = d + d(d+1)/2$ のパラメータを有し、一方、モデル M_2 は2倍のパラメータ($k_2 = 2k_1$)を有する。次式が負である場合、 i 番目のフレームがセグメント境界に対する良好な候補であることがわかる。

【数1】

$$\Delta BIC_i = -\frac{n}{2} \log|\Sigma_w| + \frac{i}{2} \log|\Sigma_f| + \frac{n-i}{2} \log|\Sigma_s| \\ + \frac{1}{2} \lambda \left(d + \frac{d(d+1)}{2} \right) \log n$$

【0035】但し、 $|\Sigma_i|$ はウインドウ全体（即ち、 n 10 個のフレームすべて）の共分散の行列式である。 $|\Sigma_f|$ はそのウインドウの第1サブディビジョンの共分散の行列式であり、 $|\Sigma_s|$ はそのウインドウの第2サブディビジョンの共分散の行列式である。

【0036】従って、ステップ610において、2つのサブサンプル（ x_1, \dots, x_i ）及び（ x_{i+1}, \dots, x_n ）が連続的なデータ・サンプル（ x_1, \dots, x_n ）のウインドウから設定される。セグメンテーション・プロセス600はステップ615乃至628において数多くのテストを行い、境界の検出があまりありそうもないロケーションにそのウインドウにおけるいくつかのBICテストが 20 対応する時、それらのテストを排除する。特に、ステップ615において、可変数 α の値が $(n/r) - 1$ の値に初期設定される。但し、 r は（フレームにおける）検出解像度である。しかる後、ステップ620において、その値 α が最大値 α_{max} を越えるかどうかを決定するためのテストが行われる。ステップ620において値 α が最大値 α_{max} を越えることが決定される場合、ステップ624において、カウンタ i が $(\alpha - \alpha_{max} + 1) \cdot r$ の値に設定される。しかし、ステップ620において、値 α が最大値 α_{max} を越えないことが決定される場合、ステ 30 ップ628において、カウンタ i が r の値に設定される。しかる後、ステップ630において、上記の式を使用してBIC値における差が計算される。

【0037】ステップ640において、カウンタ i の値が $n - r$ に等しいかどうか、換言すれば、ウインドウにおけるすべての可能なサンプルが評価されてしまったかどうかを決定するためのテストが行われる。ステップ640においてカウンタ i の値が $n - r$ に等しくないことが決定される場合、ステップ650においてその i の値が r だけインクリメントされ、ステップ630においてウインドウにおける次のサンプルに対する処理を継続する。しかし、ステップ640においてカウンタ i の値が $n - r$ に等しいことが決定される場合、ステップ660において、BIC値における最小の差（ ΔBIC_{i_0} ）が負であるかどうかを決定するための更なるテストが行われる。ステップ660において、BIC値における最小の差（ ΔBIC_{i_0} ）が負でないことが決定される場合、新しいウインドウを上記方法で考察するためにステップ610へ戻る前に、ステップ665においてウインドウ・サイズが増加させられる。従って、1つのウインドウ 50

におけるすべてのカウンタ i に対する ΔBIC 値が計算され、それらのうちのいずれも負の ΔBIC 値をもたらすものでない時、ウインドウ・サイズ n が増加させられるだけである。

【0038】しかし、ステップ660において、BIC値における最小の差が負であることが決定される場合、ステップ670において、 i_0 がセグメント境界として選択される。しかる後、ステップ675において、新しいウインドウの始まりが $i_0 + 1$ に移り、ウインドウ・サイズが N_0 に設定され、その後、新しいウインドウを上記の方法で考察するためにプログラム制御はステップ610に戻る。

【0039】従って、 i のすべての可能な値に対してBIC差のテストが行われ、最大の負の ΔBIC_i によって i_0 が選択される。そのウインドウではフレーム i においてセグメント境界が検出可能である。 $\Delta BIC_{i_0} < 0$ である場合、 x_{i_0} がセグメント境界に対応する。そのテスト結果が否定的である場合、後述のように、ステップ660において更なるデータ・サンプルが（パラメータ n を増加させることによって）現ウインドウに加えられ、フィーチャ・ベクトルがすべてセグメント化されてしまうまで、プロセスはデータ・サンプルのこの新しいウインドウに関して反復される。一般に、ウインドウ・サイズは、自身が1つのウインドウ拡張から別のウインドウ拡張に増加する複数のフィーチャ・ベクトルによって拡張される。しかし、ウインドウは、或る最大値よりも大きい多数のフィーチャ・ベクトルによっては拡張されることはない。ステップ670においてセグメント境界が検出された時、ウインドウ拡張値はそれの最小値（ N_0 ）を検索する。

【0040】E. 可変ウインドウ方式

本発明のもう1つの特徴によれば、特に小さいセグメントにおける全体の精度を改良する新しいウインドウ選択方式が提供される。セグメンテーション・プロセス600が遂行されるウインドウ・サイズの選択は非常に重要である。その選択されたウインドウがあまりにも多くのベクトルを含む場合、いくつかの境界が脱落することがある。一方、選択されたウインドウが小さ過ぎる場合、情報の不足の結果、ガウス分布によるデータの表示が不十分になるであろう。

【0041】セグメント境界が検出されなかった場合、一定量のデータを現ウインドウに加えることが提案され

た。そのような方式は、精度を改良するために「前後関係 (contextual information)」を利用するものではない。セグメント境界が検出されても又はされなくても、或いは境界が長い間検出されなくても、同じ量のデータが加えられる。

【0042】本発明の改良されたセグメンテーション・プロセスは、新しい境界が生じそうなエリアにおける比較的少量のデータを考察し、新しい境界が生じそうな時にはウインドウ・サイズをもっと大きく増加させる。まず、小さいサイズのベクトルのウインドウ（一般には、100フレームの音声）を考察する。現ウインドウにおいてセグメント境界が検出されない場合、ウインドウのサイズは ΔN_1 フレームだけ増加する。この新しいウインドウにおいて境界が検出されない場合、フレームの数は $\Delta N_{1,1}$ だけ増加する。なお、セグメント境界が検出されるまで、又はウインドウ拡張が最大サイズに達してしまうまで（境界が生じる場合に精度の問題を回避するために）、 $\Delta N_i = \Delta N_{i,1} + \delta_i$ である。但し、 $\delta_i = 2\delta_{i,1}$ である。これは、ウインドウが依然として小さい時にはかなり遅いウインドウ・サイズの増加及びウインドウが大きくなる時には速いウインドウ・サイズの増加を保証する。ウインドウ内でセグメント境界が検出される時、最小のウインドウ・サイズを使用して次のウインドウがその検出された境界の後に始まる。

【0043】F. BICテストの効率の改良

本発明のもう1つの特徴によれば、BICテストが行われるロケーションの良好な選択によって処理時間全体の改良が得られる。ウインドウにおけるBICテストのうちの或るものは、境界の検出がありそうなロケーションにそれらに対応する時、任意に排除可能である。まず、BICテストは各ウインドウの境界においては行われない。それは、それが非常にわずかなデータでもって1つのガウス分布を必ず表示するためである（この明らかに小さいゲインがセグメント検出を通して繰り返され、実際には、それは無視し得るパフォーマンス・インパクトを持たない）。

【0044】更に、現ウインドウが大きい時にBICテストがすべて行われる場合、何らかの新しい情報が加えられる度に、そのウインドウの開始時においてBIC計算が何回も行われたであろう。例えば、10秒のウインドウ・サイズにおいて最初の5秒内にセグメント境界が検出されなかった場合、10秒の現ウインドウの拡張によって、最初の5秒内に境界が認められるということは全くありそうもない。従って、（ウインドウ拡張に続く）現ウインドウの始まりにおけるBIC計算を無視することによってBIC計算の数を減少させることができる。実際には、BIC計算の最大数は、今や、必要とされる速度/精度レベルに従って調整された調節可能なパラメータ（図3における α_{xxx} ）である。

【0045】従って、セグメンテーション・プロセス6

00は、セグメンテーション情報に関する何らかのフィードバックを持つ前にそれが必要とする最大時間を知ることをする。それは、たとえセグメント境界が検出されなくても、ウインドウが十分に大きい場合、第1フレームにセグメントが存在しないということがわかるためである。この情報は速度信号のうちのこの部分に関して別の処理を行うために使用可能である。

【0046】G. BICペナルティ・ウェイト

BICの式は、理論と基準に関する実用的な応用との間の差を補うために、ペナルティ・ウェイト・パラメータ λ を利用する。ミス率と誤警報率との間の良好なトレード・オフを与える λ の最良値は1.3であることがわっている。放送ニュースの転写に対するセグメンテーション精度に関する λ の影響をより総合的に研究するためには、M.S.Thesis, Institut Eurecom 誌（フランス、1998）における A. Tritschler 氏による「BICを使用したセグメンテーション・イネーブルド音声認識アプリケーション (A Segmentation-Enabled Speech Recognition Application)」と題した論文を参照してほしい。

【0047】原則として、係数 λ はタスク依存のものであり、新しいタスク毎に戻されなければならないけれども、実際には、そのアルゴリズムは種々のタイプのデータに適用されており、同じ値の λ を使用することによるパフォーマンスにおける認め得る程度の変化は存在しない。

【0048】H. スピーカ識別プロセス

前述のように、同時転写、セグメンテーション及びスピーカ識別プロセス500は、ステップ560において、図7に示されたスピーカ識別プロセス700を実行し、登録されたスピーカ・データベース420を使用して各セグメントにスピーカ・ラベルを割り当てる。図7に示されるように、スピーカ識別プロセス700は、ステップ510において共通のフロント・エンド・プロセッサによって生成されたフィーチャ・ベクトルと共に、セグメンテーション・プロセス600によって識別されたターンを受け取る。一般に、スピーカ識別システムはスピーカ・データベース420（図4）にセグメント発声音を比較し、「最も近似した」スピーカを検出する。

【0049】ターン及びフィーチャ・ベクトルは、ステップ710において、単一のスピーカによる音声のチャックより成るセグメント発声音を形成するように処理される。ステップ720において、セグメント発声音がスピーカ識別システムに供給される。スピーカ識別システムを検討するためには、例えば、Proc. of Speaker Recognition and Its Commercial and Forensic Application, Avignon, France (1998) 誌における H.S.M.Beigi 氏他による「IBMモデル・ベース及びフレーム毎のスピーカ認識 (IBM Model-Based and Frame-By-Frame Speaker-Recognition)」と題した論文を参照してほしい。

一般に、スピーカ識別システムはセグメント発声をスピーカ・データベース420(図4)に比較し、「最も近似した」スピーカを検出する。

【0050】スピーカ識別システムは2つの異なるインプリメンテーション、即ち、モデル・ベース方式及びフレーム・ベース方式を有し、それらは付随した利点及び欠点を有する。エンジンは、放送ニュースのような番組の生のオーディオ・インデキシングを容易にするために独立したテキスト及び言語の両方である。

【0051】I. スピーカ識別(モデル・ベース方式) スピーカの母集団に対して一組のトレーニング・モデルを作成するために、下記のようなd次元のフィーチャ・ベクトルを持ったM個の音声フレームのシーケンスに基づいたi番目のスピーカに対するモデル M_i が計算される。

【数2】

$$\left\{ \vec{f}_m \right\}_{m=1, \dots, M}$$

【0052】これらのモデルは、ガウス分布が選択される場合に対して、平均ベクトル、共分散マトリックス、及びカウントより成る下記のようなそれらの統計的パラメータによって記憶される。なお、各スピーカiは n_i 個の分布よりなるモデルでもって終わり得るものである。

【数3】

$$\left\{ \vec{\mu}_{i,j}, \Sigma_{i,j}, \vec{C}_{i,j} \right\}_{j=1, \dots, n_i}$$

【0053】2つのそのようなモデルを比較するために、Proc. ICASSP98 誌(Seattle, WA, 1998)におけるH.S.M. Beigi氏他による「分布の集合体相互間の距離測定法及びスピーカ認識に対するその応用(A Distance Measure Between Collections of Distributions and Its Application to Speaker Recognition)」と題した論文において提案された距離測定法を使用して、スピーカ識別(クレームを実証する)、スピーカ分類(スピーカを割り当てる)、スピーカ検証(ラベルされたスピーカの特性に匹敵する特性を持ったスピーカの「コーホート(cohort)」セットとラベルを比較することによって分類を確認するための第2パス)、及びスピーカ・クラスタリングを含む多くの種々な機能を持ったスピーカ認識システムを考案するために階層構造が作成される。

【0054】スピーカ認識のために考案されたその距離測定法は、異なる数の分布 n_i を持った受容可能な距離の計算を可能にする。2つのスピーカをそれらのパラメ

ータ的な表示に基づいて比較するだけで、2つのスピーカを比較するというそのタスクを計算主体でないものにするという特徴を常に持つ必要がなくなる。しかし、認識段階に対するこの距離測定法の欠点は、比較の計算が始まる前に個々のテストのモデル(要求者: Claimant)を形成するために音声セグメント全体が使用されなければならないということである。フレーム・ベース方式はこの欠点を緩和する。

【0055】J. スピーカ識別(フレーム・ベース方式)

M_i をi番目の登録されたスピーカに対応するモデルであると仮定する。 M_i は、スピーカiのガウス混合モデル(GMM)の n_i 個のコンポーネントの各々に対する平均ベクトル、共分散マトリックス、及び混合ウェイトより成る次のようなパラメータ・セットによって全体的に定義される。

【数4】

$$\left\{ \vec{\mu}_{i,j}, \Sigma_{i,j}, \vec{p}_{i,j} \right\}_{j=1, \dots, n_i}$$

【0056】これらのモデルは、前のセクションにおいて説明したように、下記のようなd次元のフィーチャ・ベクトルを持ったM個の音声フレームのシーケンスより成るトレーニング・データを使用して作成される。

【数5】

$$\left\{ \vec{f}_m \right\}_{m=1, \dots, M}$$

【0057】スピーカ母集団のサイズが N_p である場合、モデル・ユニバースのセットは次のようになる。

【数6】

$$\{M_i\}_{i=1, \dots, N_p}$$

【0058】基本的な目的は、次式のようなN個のフレームのシーケンスとして表されたテスト・データを M_i が最もよく示しているというようなiを見つけること、及びそれらのモデルのうちデータを十分に記述するものがないという決定を行うことである。

【数7】

$$\left\{ \vec{f}_n \right\}_{n=1, \dots, N}$$

【0059】次のようなフレーム・ベースのウェイト付けられた距離測定法 $d_{i,n}$ はその決定を行う場合に使用される。

【数8】

$$d_{i,n} = -\log \left[\sum_{j=1}^{n_i} p_{i,j} p(f_n | j^{th} \text{ component of } M_i) \right]$$

【0060】但し、正規の表示を使用すると、次のよう * 【数9】
になる。 *

$$p(\vec{f}_n | \cdot) = \frac{1}{(2\pi)^{d/2} |\sum_{i,j} \vec{f}_n - \vec{\mu}_{i,j}|^{1/2}} e^{-\frac{1}{2}(\vec{f}_n - \vec{\mu}_{i,j})^T \sum_{i,j}^{-1} (\vec{f}_n - \vec{\mu}_{i,j})}$$

【0061】テスト・データからのモデル M_i の合計距離 D_i はテスト・フレームの合計数を越えたすべての距離の和であると見なされる。

【0062】分類のために、音声セグメントのモデルまでの最小距離を持ったモデルが選択される。その最小距離を背景モデルの距離に比較することによって、オリジナル・モードのうちのいずれも十分に合致しないことを表示するための方法を提供することが可能である。別の方法として、合計距離を計算するために投票集計技法が使用可能である。

【0063】検証のために、ラベルされたスピーカのコーホートを形成する所定セットのメンバが種々のバックグラウンド・モデルでによって増大する。このセットをモデル・ユニバースとして使用すると、テスト・データは、要求者(Claimant)のモデルが最小距離を有するかどうかをテストすることによって検証される。そうでない場合、それは拒絶される。

【0064】この距離測定法は、スピーカ相互間の距離を計算するために音声のフレームが保持されなければならないので、トレーニングでは使用されない。従って、トレーニングは、前述のモデル・ベースのテクニックのための方法を使用して行われる。

【0065】ステップ720において生成されたその割り当てられたスピーカ・ラベルは、下記のように、ユーザへ出力するために任意選択的にブロック550(図5)への暫定的提供が可能である。ステップ730において、その割り当てられたスピーカ・ラベルは、スピーカ分類の結果に関して第2パスを行うことによって検証される。ステップ730においてスピーカ識別が検証される場合、そのスピーカ・ラベルはユーザへの出力のためにブロック550(図5)に供給される。更に、ステップ740において、オリジナルの登録されたスピーカ・モデルからオーディオ・テスト・セグメントまでの距離を表す割り当てられたスコアと共に、最善の選択を表すエントリ、又は、望ましい場合には、代替えの選択を表すエントリを、スピーカ・ターン・データベース300において任意選択的に作成することが可能である。

【0066】本願において開示され及び図示された実施例並びにその変形は単に本発明の原理を説明するものであること、及び本発明の技術的範囲及び精神から逸脱することなく種々の修正を当業者が実施することが可能であることは理解されるべきである。

【0067】まとめとして、本発明の構成に関して以下の事項を開示する。

【0068】(1) 1つ又は複数のオーディオ・ソースからオーディオ情報を転写するための方法にして、前記オーディオ情報のテキスト・バージョンを作成するために前記オーディオ・ソースを転写するステップと、前記転写するステップと実質的に同時に前記オーディオ・ソースにおける潜在的なセグメント境界を識別するステップと、識別された各セグメントにスピーカ・ラベルを割り当てるステップと、を含む方法。

(2) 前記転写するステップは前記オーディオ・ソースにおける各ワードに対してタイム・アライメントを持った転写を作成することを特徴とする請求項1に記載の方法。

(3) 前記識別するステップは非同種の音声部分相互間にセグメント境界が存在するすべての可能なフレームを識別することを特徴とする請求項1に記載の方法。

(4) 前記割り当てるステップは登録されたスピーカ・データベースを利用してスピーカ・ラベルを各識別されたセグメントに割り当てることを特徴とする請求項1に記載の方法。

(5) 前記割り当てるステップは前記割り当てられたスピーカ・ラベルの信頼性を表すスコアを割り当てるステップを更に含むことを特徴とする請求項1に記載の方法。

(6) 前記割り当てるステップは前記割り当てられたスピーカ・ラベルに対して少なくとも1つの別の選択を割り当てるステップを更に含むことを特徴とする請求項1に記載の方法。

(7) 前記転写するステップ、識別するステップ、及び割り当てるステップはマルチ・スレッド環境では並列処理ブランチにおいて行われることを特徴とする請求項1

に記載の方法。

(8) 前記識別するステップはB I Cモデル選択基準を使用してセグメント境界を識別することを特徴とする請求項1に記載の方法。

(9) 1つ又は複数のオーディオ・ソースからオーディオ情報を転写するための方法にして、前記オーディオ情報からフィーチャ・ベクトルを計算するステップと、

(a) 前記オーディオ・ソースを転写して前記オーディオ情報のテキスト・バージョンを作成するために、

(b) 前記オーディオ・ソースにおける潜在的なセグメント境界を識別するために、及び(c) 各識別されたセグメントにスピーカ・ラベルを割り当てるために、前記フィーチャ・ベクトルを3つの並列処理ブランチに適用するステップと、を含む方法。

(10) 前記フィーチャ・ベクトルは共用メモリ・アーキテクチャを使用して前記並列処理ブランチに適用されることを特徴とする請求項9に記載の方法。

(11) 前記共用メモリ・アーキテクチャは前記計算されたフィーチャ・ベクトルを前記並列処理ブランチの各々に対応するチャンネルに分配することを特徴とする請求項10に記載の方法。

(12) 前記転写するステップは前記オーディオ・ソースにおける各ワードに対してタイム・アライメントを持った転写物を作成することを特徴とする請求項9に記載の方法。

(13) 前記識別するステップは非同種の音声部分相互間にセグメント境界が存在するすべての可能なフレームを識別することを特徴とする請求項9に記載の方法。

(14) 前記割り当てるステップは登録されたスピーカ・データベースを利用してスピーカ・ラベルを各識別されたセグメントに割り当てることを特徴とする請求項9に記載の方法。

(15) 前記割り当てるステップは前記割り当てられたスピーカ・ラベルの信頼性を表すスコアを割り当てるステップを更に含むことを特徴とする請求項9に記載の方法。

(16) 前記割り当てるステップは前記割り当てられたスピーカ・ラベルに対して少なくとも1つの別の選択を割り当てるステップを更に含むことを特徴とする請求項9に記載の方法。

(17) 前記識別するステップはB I Cモデル選択基準を使用してセグメント境界を識別することを特徴とする請求項9に記載の方法。

(18) 1つ又は複数のオーディオ・ソースからオーディオ情報を転写するためのシステムにして、コンピュータ読み取り可能なコードを記憶するメモリと、前記メモリに動作関係に結合され、前記コンピュータ読み取り可能なコードをインプリメントするように構成されたプロセッサと、を含み、前記コンピュータ読み取り可能なコードは、前記オーディオ・ソースを転写して前記オーディ

ィオ情報のテキスト・バージョンを作成するように、前記転写と実質的に同時に前記オーディオ・ソースにおける潜在的なセグメント境界を識別するように、及び各識別されたセグメントにスピーカ・ラベルを割り当てるように、構成されることを特徴とするシステム。

(19) コンピュータ読み取り可能なプログラム・コード手段を組み込まれたコンピュータ読み取り可能な媒体を含み、前記コンピュータ読み取り可能なプログラム・コード手段は、オーディオ情報のテキスト・バージョンを作成するためにオーディオ・ソースを転写するステップと、前記転写するステップと実質的に同時に前記オーディオ・ソースにおける潜在的なセグメント境界を識別するステップと、識別された各セグメントにスピーカ・ラベルを割り当てるステップと、を含むことを特徴とする製造物。

(20) 1つ又は複数のオーディオ・ソースからオーディオ情報を転写するためのシステムにして、コンピュータ読み取り可能なコードを記憶するメモリと、前記メモリに動作関係に結合され、前記コンピュータ読み取り可能なコードをインプリメントするように構成されたプロセッサと、を含み、前記コンピュータ読み取り可能なコードは、前記オーディオ情報からフィーチャ・ベクトルを計算し、(i) 前記オーディオ・ソースを転写して前記オーディオ情報のテキスト・バージョンを作成するために、(ii) 前記オーディオ・ソースにおける潜在的なセグメント境界を識別するために、及び(iii) 各識別されたセグメントにスピーカ・ラベルを割り当てるために、前記フィーチャ・ベクトルを3つの並列処理ブランチに適用するように構成されることを特徴とするシステム。

(21) コンピュータ読み取り可能なプログラム・コード手段を組み込まれたコンピュータ読み取り可能な媒体を含み、前記コンピュータ読み取り可能なプログラム・コード手段は、前記オーディオ情報からフィーチャ・ベクトルを計算するステップと、(i) 前記オーディオ・ソースを転写して前記オーディオ情報のテキスト・バージョンを作成するために、(ii) 前記オーディオ・ソースにおける潜在的なセグメント境界を識別するために、及び(iii) 各識別されたセグメントにスピーカ・ラベルを割り当てるために、前記フィーチャ・ベクトルを3つの並列処理ブランチに適用するステップと、を含むことを特徴とする製造物。

(22) 1つ又は複数のオーディオ・ソースからオーディオ情報を転写するための方法にして、前記オーディオ情報のテキスト・バージョンを作成するために前記オーディオ・ソースを転写するステップと、前記オーディオ・ソースにおける潜在的なセグメント境界を識別するステップと、識別された各セグメントにスピーカ・ラベルを割り当てるステップと、前記転写するステップ、識別するステップ、及び割り当てるステップと実質的に同時

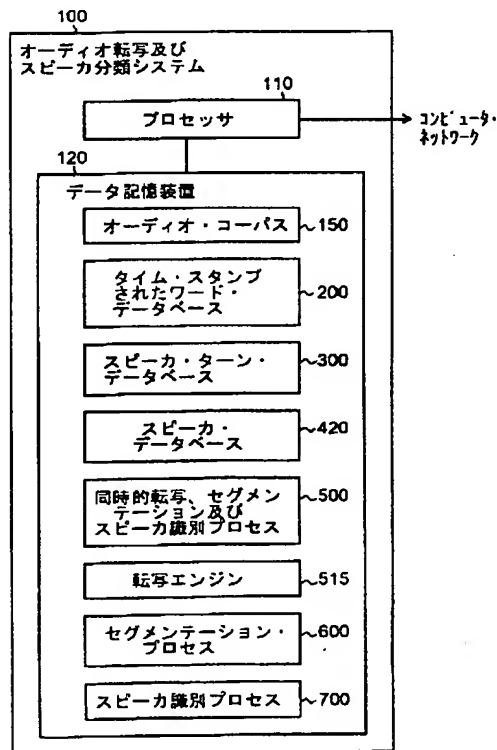
に前記テキスト・バージョンを前記割り当てられたスピーカ・ラベルと共に供給するステップと、を含む方法。

(23) 1つ又は複数のオーディオ・ソースからオーディオ情報を転写するための方法にして、前記オーディオ情報からフィーチャ・ベクトルを計算するステップと、

(i) 前記オーディオ・ソースを転写して前記オーディオ情報のテキスト・バージョンを作成するために、(i) 前記オーディオ・ソースにおける潜在的なセグメント境界を識別するために、及び(iii) 各識別されたセグメントにスピーカ・ラベルを割り当てるために、前記フィーチャ・ベクトルを3つの並列処理ブランチに適用するステップと、前記転写するステップ、識別するステップ、及び割り当てるステップと実質的に同時に前記テキスト・バージョンを前記割り当てられたスピーカ・ラベルと共に供給するステップと、を含む方法。

【図面の簡単な説明】

【図1】本発明によるオーディオ転写及びスピーカ分類*



【図1】

* システムのブロック図である。

【図2】図1のタイム・スタンプされたワード・データベースからのテーブルである。

【図3】図1のスピーカ・ターン・データベースからのテーブルである。

【図4】本発明による代表的なスピーカ登録プロセスを示す。

【図5】図1のオーディオ転写及びスピーカ分類システムによって遂行される例示的な同時転写、セグメンテーション及びスピーカ識別プロセスを説明するフローチャートである。

【図6】図1のオーディオ転写及びスピーカ分類システムによって遂行される例示的なセグメンテーション・プロセスを説明するフローチャートである。

【図7】図1のオーディオ転写及びスピーカ分類システムによって遂行される例示的なスピーカ識別プロセスを説明するフローチャートである。

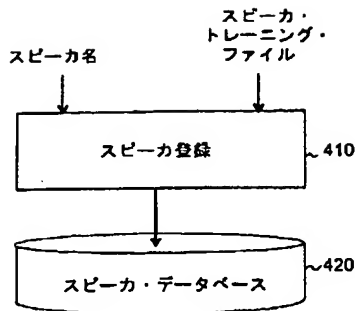
【図2】

200~ タイム・スタンプされたワード・データベース		
ワード・ストリング	開始時間	ワード期間
220	230	240
211~ 1	t ₁	
212~ 2	t ₂	
213~ ⋮	⋮	⋮
214~ N	t _N	

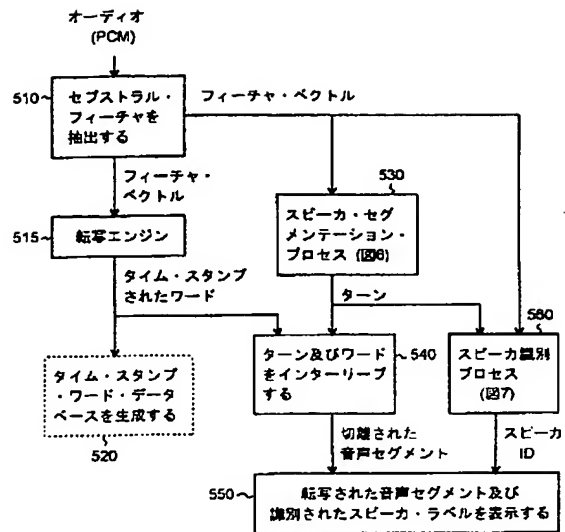
【図3】

300~ スピーカ・ターン・データベース					
セグメント番号	開始時間	スピーカ・ラベル (最良の推量)	別のスピーカ・ラベル	別のスピーカ・ラベル	別のスピーカ・ラベル
320	330	340	350	360	370
305~ 1	T _A	スピーカ X	S ₁₀	スピーカ K	S _{10-a}
306~ 2	T _K	スピーカ K	S ₁₁	スピーカ L	S _{11-a}
307~ ⋮	⋮	⋮	⋮	⋮	⋮
308~ N	T _E	スピーカ G	S ₁₂	スピーカ P	S _{12-a}

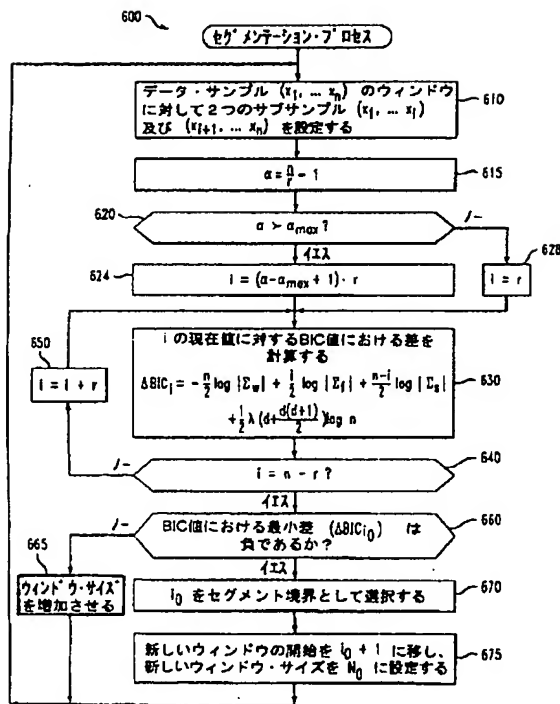
【図4】



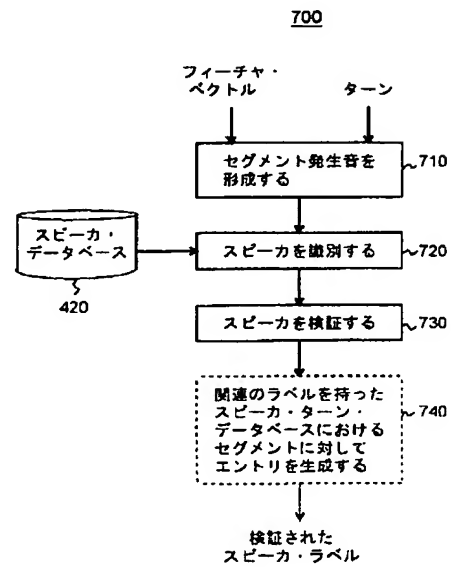
【図5】



【図6】



【図7】



フロントページの続き

(72)発明者 アラン・シャルル・ルイ・トレザー
アメリカ合衆国ニューヨーク州、ニューヨ
ーク、ウエスト・シックスティサード・ス
トリート、243 ナンバー・5・エイ

(72)発明者 マハシュ・ヴィズワナザン
アメリカ合衆国ニューヨーク州、ヨークタ
ウン・ハイツ、ダグラス・ドライブ 3024